
**D. Holt, D. G. Steel
M. Tranmer, and N. Wrigley**

Aggregation and Ecological Effects in Geographically Based Data

Statistics calculated using the means of geographic areas can differ substantially from the corresponding statistics based on data from individuals. Analysts who base their conclusions about individual-level relationships on area-level analyses run the risk of committing the ecological fallacy. Statistical models are proposed that capture the essential features of the structure of a population composed of geographically defined groups and can encompass grouping processes and contextual effects. These models are used to show how small effects in the analysis of individual-level data can be magnified substantially when the corresponding analysis based on aggregated data is carried out. Thus the source of aggregation effects is exposed. While aggregation effects have been studied by many authors, no general approach has been offered to the problem of adjusting an area-level analysis so as to correct for aggregation effects and hence remove, or at least reduce, the bias that leads to the ecological fallacy. The statistical models proposed are used to provide an approach to this problem. Data from the 1991 U.K. Census of Housing and Population are used to illustrate the size of the aggregation effects and the extent to which the proposed adjustments succeed in their objective.

1. INTRODUCTION

Statistics calculated using the means of geographic areas are often very different from those calculated directly from data on individuals. For example, the correlation coefficient between two variables calculated from the means of Census Enumeration Districts (EDs) can be quite different from the corresponding correlation coefficient calculated between the two variables measured on individuals. The same effect can apply to the regression coefficient of one variable on another. These differences are referred to as aggregation or ecological effects. If we calcu-

This research was supported by Grant no. H507 26 5013 from the Economic and Social Research Council, United Kingdom. The useful comments of the anonymous referees are also gratefully acknowledged.

D. Holt is Director, Office for National Statistics, London, and Registrar General for England and Wales. D. G. Steel is senior lecturer at the University of Wollongong. M. Tranmer is research assistant, Department of Social Statistics, University of Southampton, where N. Wrigley is professor of geography.

Geographical Analysis, Vol. 28, No. 3 (July 1996) © 1996 Ohio State University Press
Submitted 2/21/95. Final version accepted 9/5/95.

late a statistic using area means and assume that the result is an estimate of the corresponding unit-level parameter, then we run the risk of committing the ecological fallacy. Moreover, the results of the analysis of area-level data may vary according to the number of areas used and their boundaries; this is referred to as the modifiable areal unit problem (MAUP).

These effects are well known and have been investigated empirically in various studies such as those by Gehlke and Biehl (1934), Yule and Kendall (1950), Robinson (1950), Blalock (1964), Clark and Avery (1976), Openshaw and Taylor (1979), Openshaw (1984), and Fotheringham and Wong (1991). While these studies have investigated the possible effects empirically they have not provided any generally applicable theory through which the results can be interpreted or generalized. To make progress in understanding and interpreting aggregation effects we need to incorporate area effects and the way in which relationships between variables vary across areas into the statistical model upon which the analysis is based. Any model of area effects must try to account for the fact that, generally, individuals within the same area tend to be more alike than individuals in different areas.

Amrhein (1995) made a useful first step by studying aggregation effects through a simulation study in which individuals were assigned to areas and the variable values associated with them were generated randomly. This empirical study aimed to isolate the aggregation effect due to using area means from other factors, such as the fact that in practice individuals will not be assigned to areas at random but will tend to be located with others who have characteristics similar to themselves.

Steel and Holt (1996b) develop the statistical theory that applies to this "random grouping" case and obtain the statistical properties of some standard estimators, tests of hypotheses, and confidence intervals. They provide a set of rules of analysis that are applicable to this case.

However, random grouping does not usually occur in practice and more complex structures must be introduced into the statistical model to allow for the fact that individuals within any area tend to be more alike than those from different areas. Steel and Holt (1996a) identify two extensions to the statistical model that underpins the analysis, either of which would lead to individuals associated with the same area being more alike. These two models are combined into a single model which can cover a wide variety of situations. In particular:

- (i) Grouping Models—in which individuals with similar characteristics choose to live in the same area, and
- (ii) Group Dependent Models—in which individuals living in the same area are exposed to common influences and as a result exhibit similarities. This class of models includes contextual, multilevel, and variance components models.

This combined model allows a deeper understanding of how aggregation effects occur and the situations in which they will be strongest.

It is useful to understand the cause of aggregation effects but this does not solve the problem of how to adjust the results of an area-level analysis to provide reliable estimates of individual-level relationships. Duncan and Davis (1953) developed a method for calculating the possible range of a correlation coefficient from a 2×2 table with known margins. Goodman (1959) showed that ecological regression analysis could provide unbiased estimates of the corresponding individual-level parameters if the regression parameters in each area were to vary randomly about an overall value. Langbein and Lichtman (1978) considered some methods that can be applied when area membership is determined by the values of the dependent variable, and when unit-level variances are available for the

dependent variable and all the independent variables in a regression model. However, no generally applicable method of adjustment has been available.

We use the proposed combined model, together with some additional information to be specified in due course, to provide a general method of adjusting the area-level analysis to provide better estimates of relationships between variables at the individual level. We demonstrate the effectiveness of the methods proposed using U.K. Census Data.

2. MODELS FOR AREA EFFECTS

We consider a population of N individuals in some region. Associated with each individual is a set of variables of interest which we represent as a vector \mathbf{y} . The population is distributed over M areas within the region and for each individual the vector \mathbf{c}_i indicates the area to which the i th individual belongs. The number of individuals in the g th area is N_g .

We assume that over the whole population \mathbf{y} has a distribution with mean vector $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_{yy}$, which has elements σ_{ab} for the covariance between the a and b th elements of \mathbf{y} . In general, we are interested in the relationships between the variables in \mathbf{y} so our primary target for estimation is $\boldsymbol{\Sigma}_{yy}$. Once this has been considered, the theory can, in principle, be extended to the estimation of functions of $\boldsymbol{\Sigma}_{yy}$, such as the correlation matrix \mathbf{R}_{yy} , the regression coefficients of components of \mathbf{y} on others and principal components of the \mathbf{y} variables.

We assume that there is a data set consisting of a sample of n individuals in m areas for which y_i , $i = 1 \dots n$ has been observed but that these individual-level observations are unavailable to the analyst. Instead the data have been aggregated to provide a set of m vectors of area-level means, \bar{y}_g , $g = 1 \dots m$ that are available for analysis together with the sample size, n_g , upon which each vector is based. Thus the following area-level statistics can be calculated:

$$\text{The sample mean of the } g\text{th area : } \bar{y}_g \quad (1)$$

$$\text{The overall sample mean : } \bar{y} = \frac{1}{n} \sum_g n_g \bar{y}_g. \quad (2)$$

The weighted area-level sample covariance matrix:

$$\bar{S}_{yy} = \frac{1}{m-1} \sum_g n_g (\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'. \quad (3)$$

Analogous individual-level statistics may be defined but will be unavailable to the analyst. For example, the individual-level sample covariance matrix is

$$S_{yy} = \frac{1}{n-1} \sum_i (y_i - \bar{y})(y_i - \bar{y})'. \quad (4)$$

Throughout this paper it is assumed that the sample design can depend on the areas but not on \mathbf{y} or any variable which is related to \mathbf{y} , given area membership. For example, a census or a simple random sample of areas and individuals within areas may be used.

2.1 Random Composition of Areas

Amrhein (1996) simplified the usual situation by investigating empirically the effect of using area means when the allocation of individuals to areas is completely random. Steel and Holt (1996b) provide the statistical theory for this situation and obtain the properties of statistics such as means, variances, and regression and correlation coefficients. This leads to the properties of aggregated statistics calculated from randomly formed areas and rules for the analysis of such data.

For a member of the g th area the underlying model assumption that is consistent with randomly formed areas is that

$$y_i = \mu_y + \varepsilon_i \quad \text{for } i \in g \text{ (that is, } c_i = g) \tag{5}$$

where the components of the vector of individual deviations ε_i may be correlated, but for different individuals these random terms are independent, irrespective of the area to which they belong.

Let $\Sigma_{\varepsilon\varepsilon}$ be the covariance matrix of the individual deviations; then

$$V(\varepsilon_i) = \Sigma_{\varepsilon\varepsilon} = \Sigma_{yy};$$

$$\text{Cov}(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \text{for all } i \neq i'.$$

For common statistics such as means, variances, and regression and correlation coefficients, calculated using area sample sizes as weights, the expectation is not affected by aggregation and we continue to obtain unbiased estimates of the corresponding population parameters. In particular:

$$E(\bar{y}) = \mu_y, \quad \text{and} \quad E(\bar{S}_{yy}) = \Sigma_{yy}$$

However, there is an aggregation effect since the variation of some statistics is affected in a way that is mainly related to the number of areas used in the analysis. The usual number of degrees of freedom will be reduced and confidence intervals will be wider than if individual data had been used. Hence, procedures such as tests of hypotheses and estimation of confidence intervals must be modified.

To illustrate some of the effects, consider the mean calculated from area data,

$$\bar{y} = \frac{1}{n} \sum_g n_g \bar{y}_g.$$

This is arithmetically identical to the sample mean of the original individual observations. The fact that it is based on the area means has not led to any loss of information at all and there is no aggregation effect whatsoever. Hence,

$$V(\bar{y}) = \frac{\Sigma_{yy}}{n}$$

which is precisely the variance that would have been achieved if individual observations had been used.

However, the same is not true for \bar{S}_{yy} since, to a first approximation, $V(\bar{S}_{yy})$ is n/m times as large as $V(S_{yy})$, the corresponding sample covariance based on individual observations. Similar effects apply to regression and correlation coefficient derived from \bar{S}_{yy} .

Even for inferences about μ_y we must be careful. The point estimate \bar{y} for μ_y is unbiased and has the same variance we would obtain from individual data. However, if we wish to estimate a $(1 - \alpha)$ 100 percent confidence interval for one element of \bar{y} , say \bar{y}_a , based on individual-level data, the standard method would be to use

$$\bar{y}_a \pm t_{\alpha/2, n-1} \sqrt{\frac{s_{aa}}{n}}$$

where s_{aa} is the (a, a) th element of S_{yy} and is the individual-level variance for y_a . $t_{\alpha, n-1}$ is the appropriate $(1 - \alpha)$ 100 percent point of the t distribution with $n - 1$ degrees of freedom. If we use the corresponding area-level statistic \bar{s}_{aa} , then the appropriate confidence interval is

$$\bar{y}_a \pm t_{\alpha/2, m-1} \sqrt{\frac{\bar{s}_{aa}}{n}}$$

Even though the mean and variance of \bar{y}_a are the same whether individual or aggregated data are used, the confidence interval is widened by using m , the number of areas, rather than n , the number of observations, to determine the degrees of freedom of the relevant t distribution. Note that for \bar{y}_a the only difference in the confidence interval is the use of m , rather than n , for the degrees of freedom for the t distribution. When m is large (that is, greater than forty) this will not be important. However, this illustrates that care is needed when analyzing aggregated data. For other statistics such as regression coefficients the impact can be much greater.

With proper allowance for the variation associated with the area-level analyses, Steel and Holt (1996b) provide methods for point estimation, estimation of standard errors and confidence intervals, and modifications to tests of hypotheses for randomly composed areas.

2.2 Area Effects

While the case of randomly allocating individuals to areas is informative and can strengthen our understanding, it is not a situation that usually occurs in practice. It is well known that two individuals who live in the same area have characteristics in common and tend to be more alike in terms of a wide range of socioeconomic and health-related variables than two individuals drawn at random from the whole population. Thus the group of individuals who live in an area are more homogeneous than the population as a whole. This phenomenon has been observed in surveys employing cluster or multistage sampling for many years, where the areas are described as exhibiting a positive intracluster correlation.

A simple way to represent this positive clustering is through a variance-components model in which area-level random effects are introduced to allow for positive intra-area correlation. Thus (5) is extended as follows:

$$y_i = \mu_y + v_g + \varepsilon_i \quad \text{for } i \in g. \quad (6)$$

Here v_g is a vector of unobserved area-level effects that vary randomly between groups (one component of the vector corresponding to each variable of interest). The random effects in v_g may be correlated across variables within the same area but the vectors of random effects for two different areas are assumed to be independent;

$$\begin{aligned} \text{Cov}(\mathbf{v}_g, \mathbf{v}_{g'}) &= \Delta_{yy} & g = g' \\ &= 0 & \text{otherwise.} \end{aligned}$$

In matrix notation this model may be written as follows:

Model A:

$$E(y_i | c) = \boldsymbol{\mu}_y. \tag{7}$$

$$V(y_i | c) = \boldsymbol{\Sigma}_{yy}. \tag{8}$$

$$\begin{aligned} \text{Cov}(y_i, y_j | c) &= \Delta_{yy} & \text{if } c_i = c_j & \quad i \neq j \\ &= 0 & \text{otherwise.} \end{aligned} \tag{9}$$

The notation $V(\cdot | c)$ implies the covariance matrix conditional on the group labels c but unconditional over group-level random effects. Thus $\boldsymbol{\Sigma}_{yy}$ contains the within-group covariance matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \boldsymbol{\Sigma}_{yy} - \Delta_{yy}$ plus the area-level covariance matrix Δ_{yy} .

This model allows for spatial autocorrelation at the individual level, in which the values of individuals within the same area are equally correlated due to the Δ_{yy} term. Each small area is treated as a neighborhood within which individuals are correlated. Each individual in group g is “connected” with each of the other $n_g - 1$. In practice there may be correlations between individuals in different areas, such as adjacent areas. However, the model proposed here will be a useful first approximation to the more complex correlation structure that may apply. Furthermore, in the next section we propose a way of accounting for the between-individual correlations through certain auxiliary variables. Once this is done effectively there should be less between-individual correlation.

Steel and Holt (1996a) show that, for the sample selected s , the properties of the individual- and area-level statistics under Model A, for the groups formed, are

$$E[\bar{y} | c] = \boldsymbol{\mu}_y \tag{10}$$

$$E[S_{yy} | c] = \boldsymbol{\Sigma}_{yy} - \frac{\bar{n}^0 - 1}{n - 1} \Delta_{yy} \tag{11}$$

$$E[\bar{S}_{yy} | c] = \boldsymbol{\Sigma}_{yy} + (\bar{n}^* - 1) \Delta_{yy} \tag{12}$$

where $\bar{n} = n/m$, $\bar{n}^0 = \frac{1}{n} \sum_g n_g^2 = \bar{n}(1 + c_n^2)$, $\bar{n}^* = \bar{n} \left(1 - \frac{c_n^2}{m-1}\right)$ and $c_n^2 = \frac{1}{m} \boldsymbol{\Sigma}_g (n_g - \bar{n})^2 / \bar{n}^2$ is the square of the coefficient of variation of the group sample sizes in the sample.

These results show the effect of introducing components of variance to allow for area effects. The aggregate estimate of the mean is still unbiased for $\boldsymbol{\mu}_y$ but the estimator of variance is affected. The individual-level sample covariance matrix S_{yy} cannot, of course, be calculated from the area means, but we note that it is biased for $\boldsymbol{\Sigma}_{yy}$ by a term determined by Δ_{yy} . However, typically the components of Δ_{yy} are much smaller than those of $\boldsymbol{\Sigma}_{yy}$. From sample survey experience, if the components of Δ_{yy} are as much as one-tenth of those of $\boldsymbol{\Sigma}_{yy}$, this would be a very strong area effect. Furthermore, the coefficient of Δ_{yy} in (11) is approximately $1/m$ so that if a large number of areas are used (say, a hundred) and a component of Δ_{yy} is one-tenth of the corresponding term in $\boldsymbol{\Sigma}_{yy}$, then the bias will be very small indeed, that is, a bias of 0.1 percent of $\boldsymbol{\Sigma}_{yy}$.

However, when we consider the aggregate analysis the picture is changed considerably. The bias of \bar{S}_{yy} in (12) is also a multiple of Δ_{yy} but the coefficient is approximately \bar{n} . Hence, if a component of Δ_{yy} is one-tenth of the corresponding component of Σ_{yy} , an average area sample size of ten will lead to a bias that is as large as the term we are trying to estimate. This illustrates how a small bias for the analysis based on individual observations can be magnified into a much larger bias in the area-level analysis and of a different sign. This is the source of the aggregation effect and hence the ecological fallacy. Many analyses of geographically aggregated data are effectively based on a complete census and the average population of areas can often be hundreds or thousands of people, leading to very large aggregation effects. The model introduced here has important implications for the MAUP, which are discussed by Holt, Steel, and Tranmer (1996).

2.3 Area Composition Models

The previous section shows that when we make allowance for area homogeneity through a set of random effects for areas, the effect on the area-level analysis can be very great. The bias that is introduced will distort the area-level analysis and lead to aggregation effects and the ecological fallacy. Essentially the between-area differences, and hence the within-area homogeneity, is drawn in to the area-level analysis and confounded with the individual-level effects.

In the discussion of ecological analysis, models have been proposed that take into account the area formation process. In such an approach it is assumed that there is a grouping process that allocates individuals to areas according to a vector of grouping or auxiliary variables, z_i , either stochastically or deterministically. This approach is implicit in Blalock's (1964) analysis and used explicitly by Hannan and Burstein (1974), Lichtman (1974), Langbein and Lichtman (1978), Smith (1977), and Blalock (1979, 1985).

We propose in this section an explicit formulation of the relationship between the variables of interest and the grouping variables while still making provision for residual area homogeneity.

In such models it is assumed that area membership arises by some process involving the grouping variables that are associated with the variables of interest. The multivariate version of this model that combines both the effect of the grouping variables, z_i , and random effects for residual within-area correlation is Model B

$$E(y_i | z, c) = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \quad (13)$$

$$V(y_i | z, c) = \Sigma_{yy.z} \quad (14)$$

$$\begin{aligned} \text{Cov}(y_i, y_j | z, c) &= \Delta_{yy.z} && \text{if } c_i = c_j && i \neq j \\ &= 0 && \text{otherwise.} \end{aligned} \quad (15)$$

This model allows for area formation processes that are characterized by the auxiliary variables z_i . The same model can take account of contextual variables that influence individuals who live in a specific area by including them as components of z . Variables used in the sample design, for example, stratification variables, may also be included in z . The area effects of Model A, Δ_{yy} are now replaced by residual within-area covariances $\Delta_{yy.z}$ that reflect random effects after allowing for the grouping variables, z_i . Hence, this model combines grouping- and area-level

effects into a single combined model. This model seeks to at least partly explain the within-area correlations observed in the y variables through the within-area correlations of the auxiliary variables.

For the sample selected, the properties of the individual- and area-level statistics under Model B are (Steel and Holt 1994) as follows:

$$E[\bar{y} | z, c] = \boldsymbol{\mu}_y + \beta'_{yz}(\bar{z} - \mu_z) \tag{16}$$

$$E[S_{yy} | z, c] = \boldsymbol{\Sigma}_{yy} + \beta'_{yz}(S_{zz} - \boldsymbol{\Sigma}_{zz})\beta_{yz} - \frac{\bar{n}^0 - 1}{n - 1} \Delta_{yy.z} \tag{17}$$

$$E[\bar{S}_{yy} | z, c] = \boldsymbol{\Sigma}_{yy} + \beta'_{yz}(\bar{S}_{zz} + \boldsymbol{\Sigma}_{zz})\beta_{yz} + (\bar{n}^* - 1) \Delta_{yy.z} \tag{18}$$

Equations (17) and (18) show the effect of introducing grouping variables to explain some of the between-area differences. If we consider S_{yy} , for example, the bias shown in Model A of

$$-\frac{\bar{n}^0 - 1}{n - 1} \Delta_{yy} \tag{19}$$

has been partitioned into two components

$$\beta'_{yz}(S_{zz} - \boldsymbol{\Sigma}_{zz})\beta_{yz} \tag{20}$$

and

$$-\frac{(\bar{n}^0 - 1)}{n - 1} \Delta_{yy.z}. \tag{21}$$

The first term depends on the difference between the sample covariance matrix S_{zz} for the grouping variables and the population covariance matrix $\boldsymbol{\Sigma}_{zz}$. This term also depends on the strength of the relationship between the variables of interest, y , and the grouping variables, z , as determined by the matrix of regression coefficients of y on z , β_{yz} .

The second term in the bias of S_{yy} is exactly the same as in Model A except that the covariance matrix of area-level effects Δ_{yy} has been replaced by the covariance matrix of residual area effects $\Delta_{yy.z}$. If z is successful in explaining much of the between-area differences of the y variables, then the elements of $\Delta_{yy.z}$ will be much smaller than those of Δ_{yy} and this second term will be much smaller than the bias term in Model A. The difference between the bias under Model A and the residual bias under Model B will be taken up by the first component of bias (20).

Equation (18) shows that the aggregation bias can be partitioned into two components. The first component depends on the aggregation effect for the auxiliary variables, which can be considerable. The second component is a multiple of the residual area effects, $\Delta_{yy.z}$, with coefficient of order \bar{n} .

The weighted group-level matrix \bar{S}_{yy} is intended to estimate $\boldsymbol{\Sigma}_{yy}$. The first bias term in (18) is due to the effect of the grouping variables and will be zero if $\beta_{yz} = 0$ or approximately so if \bar{S}_{zz} is approximately equal to $\boldsymbol{\Sigma}_{zz}$. These conditions are quite strong and unlikely to apply in practice so that this first term is likely to contribute substantially to the bias of \bar{S}_{yy} in many situations. The effect of aggregation on the random effects for Model B is similar to Model A since the

coefficient of $\Delta_{yy,z}$ is approximately \bar{n} . However, we expect that the effect of z is to reduce the size of elements of $\Delta_{yy,z}$ compared with Δ_{yy} and thus reduce this component of the bias. Hence, introduction of the auxiliary variables enables us to explain at least some of the aggregation effects.

3. AN EMPIRICAL ANALYSIS OF AGGREGATION EFFECTS

To show how these models enable the aggregation effects to be investigated and decomposed, we consider an analysis of the 1991 U.K. population census data for the Local Authority District (LAD) of Reigate, Banstead, and Tandridge which is in Surrey, south of London. The district contains approximately 188,700 people living in 371 Census Enumeration Districts (EDs), giving an average number of people per ED of $\bar{n} = 508.63$. These EDs are taken to be the geographic areas for the analysis. For this LAD the coefficient of variation of the ED population sizes, n_g , is $C_n = 0.227$ and since $m = 371$ is large, $\bar{n}^* = 508.56$ is effectively the same as \bar{n} . Area-level data are available from the census on a complete count basis for each ED in the LAD. Corresponding unit-level data for the LAD, but not for individual EDs, are obtained from a 2 percent sample of anonymized records (SAR) released from the census as a public use sample. Hence the SAR contains approximately 3,700 individual records for this region. The following analysis is based on seventeen census variables observed for each person.

3.1 No Auxiliary Variables

For each variable, a , the area-level data were used to calculate the area-level variance \bar{s}_{aa} , and the individual-level (SAR) data were used to calculate the individual-level variance s_{aa} . A simple measure of the aggregation effect for variable a is obtained using $\bar{Q}_{aa} = \bar{s}_{aa}/s_{aa}$, the ratio of the area-level and individual-level variance. The aggregation bias is $\bar{s}_{aa} - s_{aa}$. Based on (12) we can also calculate an estimate, $\hat{\Delta}_{aa}$, of the element of Δ_{yy} corresponding to variable a and hence the bias due to aggregation $(\bar{n}^* - 1)\hat{\Delta}_{aa}$. Furthermore, it is easy to convert these values into an estimate of the intra-area correlation, $\hat{\delta}_{aa} = \hat{\Delta}_{aa}/\sigma_{aa}$, of each variable. The intra-area correlation is the correlation of values of the variable a between different individuals in the same area and is a measure of the within-area homogeneity commonly used in sample surveys (see Hansen, Hurwitz, and Madow 1953). The estimated intra-area correlation is

$$\hat{\delta}_{aa} = \frac{\bar{s}_{aa} - s_{aa}}{(\bar{n}^* - 1)s_{aa}} = \frac{\hat{\Delta}_{aa}}{s_{aa}}. \quad (22)$$

Table 1 summarizes the effect of aggregation for these seventeen variables included in the study by giving the estimated unit-level variance, s_{aa} , and area-level variances, \bar{s}_{aa} , the aggregation effect \bar{Q}_{aa} , and the estimated intra-area correlation, $\hat{\delta}_{aa}$.

From Table 1 we see that the intra-area correlations are generally small, with all values below 0.07, and most less than 0.02. The median value is 0.012. From practical experience of sample surveys, a value of 0.1 represents a strong area effect for most socioeconomic variables. However, the impact of $(\bar{n}^* - 1)$ when $\bar{n}^* - 1 = 507$ is very strong and the resulting aggregation effects and the bias of \bar{s}_{aa} are extremely large. This illustrates numerically the theoretical consequence that large aggregation bias can occur even when area effects are small, if the area means are based upon large numbers of observations. The aggregation

TABLE 1
Aggregation Effects for Variances of Variables of Interest

Variable	Unit-Level Variance	Area-Level Variance	Aggregation Effect	Intra-area Correlation
Female	0.25	0.27	1.08	0.0002
Unemployed	0.03	0.06	2.27	0.0025
HoH born N.C. [†]	0.02	0.08	3.59	0.0051
Student ≥ 18	0.02	0.08	4.17	0.0062
HoH born U.K.	0.06	0.27	4.48	0.0068
Aged 30–44	0.18	0.82	4.56	0.0070
Aged 45–59*	0.15	0.90	5.97	0.0098
Married	0.25	1.56	6.24	0.0103
Long-term illness	0.08	0.58	7.24	0.0123
Non-white*	0.03	0.22	8.27	0.0143
Fulltime Employment	0.22	1.88	8.55	0.0149
Migrant HoH	0.09	0.82	9.04	0.0158
Aged 18–29	0.13	1.18	9.20	0.0162
Other employment status	0.19	2.14	11.20	0.0201
Aged 60 and over*	0.17	2.87	17.17	0.0319
≤ 0.5 persons/room	0.25	6.93	27.96	0.0531
0 car households	0.10	3.19	32.98	0.0630

NOTES: * Chosen as auxiliary variables. [†] HoH is an abbreviation for Head of Household. NC is an abbreviation for New Commonwealth. Source: Derived from 1991 Census SAS and SAR data; Crown Copyright.

effects vary from only 1.08 for “Females” to 33 for the variable “0 Car Household.” The median aggregation effect is 7.24.

To understand the effects of aggregation on statistics such as regression and correlation coefficients we must also examine the effect of aggregation on covariances. A similar approach may be applied to every combination of two variables of interest. We may identify the aggregation effects as the ratio of the area-level estimates of covariances to the individual-level estimates from the SAR, that is, $\bar{Q}_{ab} = \bar{s}_{ab}/s_{ab}$. The bias due to aggregation is given by $(\bar{n}^* - 1) \Delta_{ab}$, where Δ_{ab} is the ab th element of Δ_{yy} and is estimated by $\hat{\Delta}_{ab} = (\bar{s}_{ab} - s_{ab})/(\bar{n}^* - 1)$. Since there are seventeen variables of interest there are 136 combinations of two different variables and a tabular presentation of the aggregation effects and area components of variance is inappropriate. The estimated intra-area cross-correlation, analogous to $\hat{\delta}_{aa}$, is given by

$$\hat{\delta}_{ab} = \frac{\hat{\Delta}_{ab}}{\sqrt{s_{aa}s_{bb}}}$$

which estimates the intra-area correlation for variables a and b . Figure 1b shows a histogram of the estimated intra-area cross-correlations $\hat{\delta}_{ab}$; for comparison Figure 1a shows the corresponding histogram of the intra-area correlations $\hat{\delta}_{aa}$ given in Table 1. We note that in general the values of $\hat{\delta}_{ab}$ are smaller than those for $\hat{\delta}_{aa}$ and most of the former fall in the range -0.01 to 0.01 . We can express the aggregation effect for a covariance as

$$\bar{Q}_{ab} = 1 + (\bar{n}^* - 1)\hat{\delta}_{ab}/r_{ab}$$

where r_{ab} is the individual-level correlation between variables a and b . Figure 2b contains a histogram of the aggregation effects for covariances and Figure 2a contains the aggregation effects for variances from Table 1.

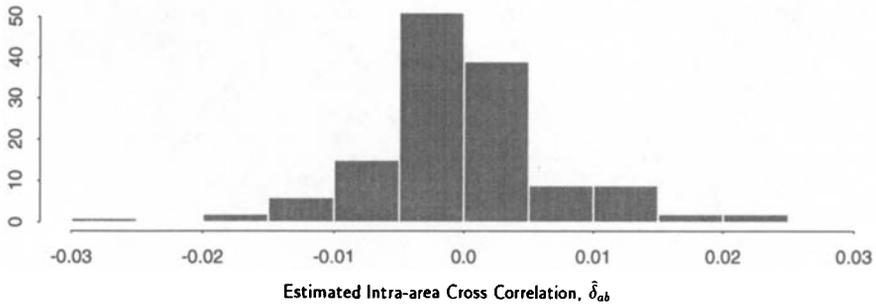
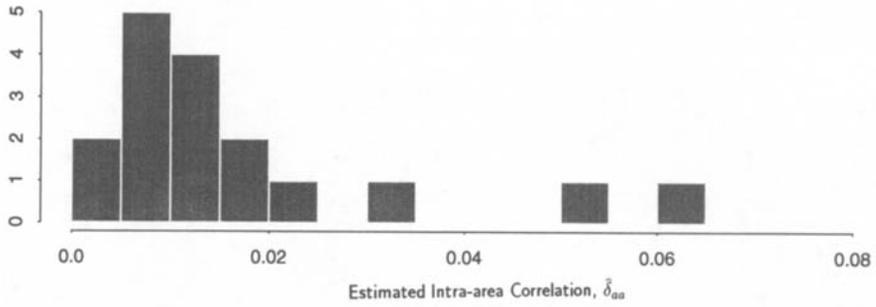


FIG. 1. Fig. 1a (top): Histogram of Intra-area Correlations. Fig. 1b (bottom): Histogram of Intra-area Cross-correlations.

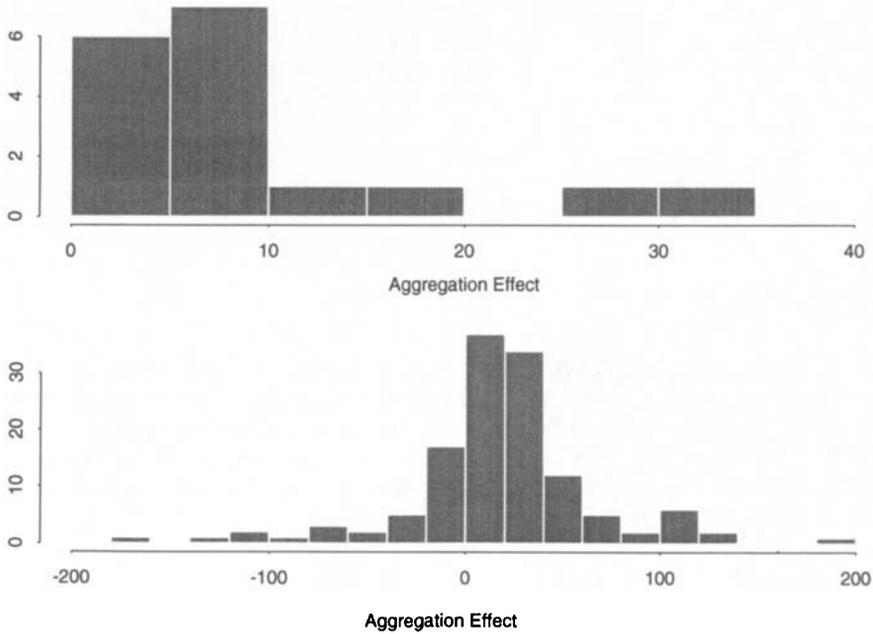


FIG. 2. Fig. 2a (top): Histogram of Aggregation Effects on Variances. Fig. 2b (bottom): Histogram of Aggregation Effects on Covariances.

Because of the impact of $(\bar{n}^* - 1)$ and r_{ab} the resulting aggregation effect can be considerable even though the values of $\hat{\delta}_{ab}$ are very small. The median aggregation effect on covariances is 19.3 although the range covers high positive and negative values. Some extreme values of the ratio are obtained due to small unit-level covariances. If r_{ab} is close to zero then \bar{Q}_{ab} can be extremely large in absolute value. Generally, however, the aggregation effect is between 0 and 80. Five values outside the range -200 to 200 are not shown and correspond to very small individual-level covariances.

3.2 Introducing Auxiliary Variables

Model B allows for the possibility that some of the area effects may be explained by a set of auxiliary variables z . As a result, the aggregation bias estimated under Model A can be partitioned into two components: one related to the auxiliary variables and the other determined by residual intra-area effects, after allowing for the effects of the auxiliary variables.

To illustrate the approach we identify a set of basic demographic and housing variables for use as auxiliary variables. Some of these are contained in the seventeen variables of interest, namely, the age categories 45–59 and 60 and over, and nonwhite (as indicated in Table 1). The other variables are characteristics of housing and are not contained in the variables of interest. These are listed in Table 2 together with the aggregation effects on their variances and this demonstrates the much higher within-area homogeneity found for housing characteristics. The housing variables are known to be strongly related to a wide variety of socioeconomic variables in the United Kingdom and therefore should be valuable auxiliary variables.

Using the partition of bias expressed in (18) we can estimate the extent to which the original aggregation bias can be attributed to the auxiliary variables z , that is, $\beta'_{yz} (\bar{S}_{zz} - \Sigma_{zz}) \beta_{yz}$, and that which remains in the residual term $(\bar{n}^* - 1) \Delta_{yy.z}$. Let τ_{ab} be the ratio of the bias due to the auxiliary variables to the original bias. That is τ_{ab} is the (ab) th element of $\beta'_{yz} (\bar{S}_{zz} - \Sigma_{zz}) \beta_{yz}$ divided by $(\bar{n}^* - 1) \Delta_{ab}$.

Table 3 contains the area-level variances, \bar{s}_{aa} , together with the unit-level variances, s_{aa} , the component of bias due to the chosen auxiliary variables, and the residual bias term. The final three columns sum to the first. In many cases the bias due to the auxiliary variables is a large component of the total bias. Excluding the three variables chosen as auxiliary variables, the median percentage of the total bias attributed to the auxiliary variables is 75 percent. For the three variables chosen as auxiliary variables, this component accounts for 100 percent of the bias and the residual bias is zero. Notice that for “Females,” for which the bias is very small, the bias attributed to the auxiliary variables exceeds the

TABLE 2
Aggregation Effects for Housing Variables

Variable	Unit-Level Variance	Area-Level Variance	Aggregation Effect	Intra-area Correlation
Tenure:				
Owned*	0.16	14.44	90.83	0.1770
Local Authority Rented*	0.11	14.25	133.43	0.2609
Stock:				
Amenities*	0.11	6.26	58.52	0.1133
Type*	0.12	10.44	90.03	0.1754

NOTES: * Chosen as auxiliary variables.

TABLE 3
Decomposition of Aggregation Bias for Variances

Variable	Area-Level Variance	Unit-Level Variance	Bias due to Z	Residual Bias
Female	0.27	0.25	0.10	-0.08
Unemployed	0.06	0.03	0.02	0.01
HoH born N.C.	0.08	0.02	0.05	0.01
Student ≥ 18	0.08	0.02	0.03	0.03
HoH born U.K.	0.27	0.06	0.12	0.09
Aged 30-44	0.82	0.18	0.60	0.05
Aged 45-59*	0.90	0.15	0.75	0.00
Married	1.56	0.25	1.10	0.21
Long-term illness	0.58	0.08	0.40	0.10
Nonwhite*	0.22	0.03	0.19	0.00
Fulltime Employment	1.88	0.22	0.73	0.93
Migrant HoH	0.82	0.09	0.21	0.52
Aged 18-29	1.18	0.13	0.65	0.40
Other employment status	2.14	0.19	1.83	0.13
Aged 60 and over*	2.87	0.17	2.71	0.00
≤ 0.5 persons/room	6.93	0.25	4.19	2.50
0 car households	3.19	0.10	2.58	0.51

Footnotes to Table 1 apply.

Source: Derived from 1991 Census SAS and SAR data; Crown Copyright.

original bias and hence a negative residual bias is estimated. This is consistent with statistical theory for variables such as this since the proportion of females in each area may be less variable than one would expect if individuals were allocated to areas completely randomly without regard to sex.

The same partition of the aggregation bias can be provided for the covariance terms although there are too many to report as a tabulation. Figure 3b shows a histogram of τ_{ab} , the ratio of the bias attributed to the auxiliary variables to the original bias. Figure 3a shows the corresponding histogram of τ_{aa} for the variances presented in Table 3. In Figure 3b we note that for a small number of cases the ratio is greater than one. This arises if the bias term due to the auxiliary variables is larger in absolute value than the original bias and will leave a residual bias of opposite sign. However, the residual bias will be smaller in absolute value than the original bias for all cases where the ratio shown is less than 2. The median ratio of the bias due to the auxiliary variables compared with the original bias is 0.94, showing that the auxiliary variables account for a very significant component of the original aggregation bias in many cases. In almost all cases the auxiliary variables account for over 60 percent of the original bias.

It follows from these results that a substantial proportion of the aggregation bias would be removed if we could take account of the bias due to the auxiliary variables as part of the estimation process. This approach is developed in the next section.

Moellering and Tobler (1972) used analysis of variance techniques to partition the total variation between the lowest level of geographic areas into components attributable to various scale in situations where they had a nested hierarchical geographic data structure. For the case of two levels their approaches is based on a statistical model equivalent to equation (6) in which v_g and ε_g are treated as fixed effects. They consider the single variable case and use unit-level data with geographic indicatives to estimate the variance components. By treating the components v_g and ε_g as random effects the variance components can be given a useful interpretation in terms of intra-area correlations. Our approach is multivariate and seeks to attribute some of the variance components,

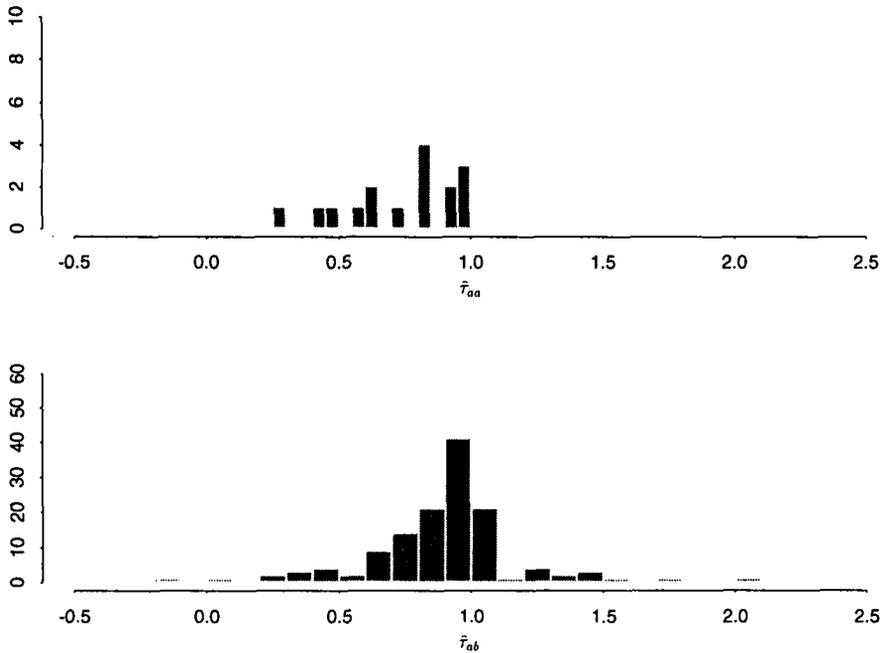


FIG. 3. Fig. 3a (top): Bias due to *Z* as a Ratio of Original Bias: Variances. Fig. 3b (bottom): Bias due to *Z* as a Ratio of Original Bias: Covariances.

and hence the causes of the aggregation effects, to the auxiliary variables. Moreover, the estimation method proposed here does not require unit-level data with small area indicatives, but only data that allows the overall unit-level covariance matrix S_{yy} to be calculated.

4. ADJUSTMENT OF AGGREGATION BIAS

Model A provides a framework through which we may understand the factors which influence aggregation bias and demonstrate how small effects that exist at the individual level are magnified through the area-level analysis to result in substantial bias effects. Model B introduces the idea of a set of auxiliary or grouping variables, which account for a substantial proportion of the aggregation bias. These results lead to a possibility, not previously suggested, that in certain circumstances, we may adjust an area-level analysis to remove a substantial proportion of the aggregation bias (that is, that attributable to *z*). This would result in an ecological analysis that is a better estimation procedure for the target of inference Σ_{yy} .

The aggregation bias attributed to the auxiliary variables, *z*, is given in (18):

$$\beta'_{yz} (\bar{S}_{zz} - \Sigma_{zz}) \beta_{yz}. \tag{23}$$

If we could obtain an estimate of this component, then it could be subtracted from \bar{S}_{yy} and so remove a major component of bias.

From Model B it may be shown that we may obtain an unbiased estimate of β_{yz} from the area-level data so long as the area means for *z* are available as well as those for *y*. Steel and Holt (1996a) show that the area-level estimator

$$\bar{B}_{yz} = \bar{S}_{zz}^{-1} \bar{S}_{yz}$$

is an unbiased estimator for β_{yz} .

Since \bar{S}_{zz} is available from the area-level data, the only remaining unknown element of (23) is Σ_{zz} . This is the individual-level covariance matrix of the z variables.

Suppose that the set of auxiliary variables are such that an estimate $\hat{\Sigma}_{zz}$ of the individual-level covariance matrix Σ_{zz} for the LAD as a whole is available from some source. This source may be quite separate from the data used in the rest of the analysis.

In our illustration, for example, we chose for z the basic demographic variables and housing-related variables described in section 3.2. Whatever the set of variables of interest, there will often exist variables for which an estimate $\hat{\Sigma}_{zz}$ of the individual-level covariance matrix Σ_{zz} can be obtained from census data or some other published source. We may then obtain an estimate of this component of the aggregation bias by using

$$\bar{B}'_{yz}(\bar{S}_{zz} - \hat{\Sigma}_{zz})\bar{B}_{yz}.$$

Hence,

$$\tilde{\Sigma}_{yy} = \bar{S}_{yy} + \bar{B}'_{yz}(\hat{\Sigma}_{zz} - \bar{S}_{zz})\bar{B}_{yz} \quad (24)$$

will be an estimator for Σ_{yy} that has residual bias $(\bar{n}^* - 1)\Delta_{yy,z}$ and will be a substantial improvement on the unadjusted area-level analysis. The residual bias will be greatly reduced if z can be chosen to explain as much as possible of the aggregation bias and hence reduce $\Delta_{yy,z}$ as closely as possible to zero.

We illustrate how this approach works by returning to the previous numerical example and the auxiliary variables discussed there. From the original area-level analysis and the corresponding individual-level analysis we may obtain the correlation between each pair of variables of interest. We denote these r_{ab} and \bar{r}_{ab} , respectively.

Figure 4a shows a plot of the area-level correlations, \bar{r}_{ab} , versus the individual-level correlations, r_{ab} , without taking any account of the auxiliary variables. If there were no aggregation effects, the points would be clustered around the line $\bar{r}_{ab} = r_{ab}$ shown in Figure 4a but we see instead a characteristic S-shaped plot. The plot shows that in general the area-level correlations \bar{r}_{ab} are of the same sign but larger in absolute value than the individual-level correlations r_{ab} .

If we take into account the auxiliary variables z , then we may adjust the area-level analysis by using $\tilde{\Sigma}_{yy}$ in (24) to calculate the adjusted correlation \tilde{r}_{ab} . Figure 4b shows the plot corresponding to Figure 4a except that the adjusted correlations \tilde{r}_{ab} are plotted against r_{ab} . We note that the size of the aggregation bias has been greatly reduced and the S-shaped plot has been replaced by a cloud of points that essentially follow the line $\tilde{r}_{ab} = r_{ab}$. In fact, the deviations away from this line are generally consistent with the size of sampling variation that would be observed from correlations based on $m = 371$ observations.

Our conclusion is that, for this example at least, the auxiliary variables are extremely successful at removing the aggregation bias and the results of the adjusted analysis will not be as misleading as results based on the unadjusted analysis. Conclusions based on the adjusted analysis will significantly reduce the impact of the ecological fallacy.

In practice the identification of an effective set of auxiliary variables is a key part of the proposed methodology. Several approaches can be used (see Steel,

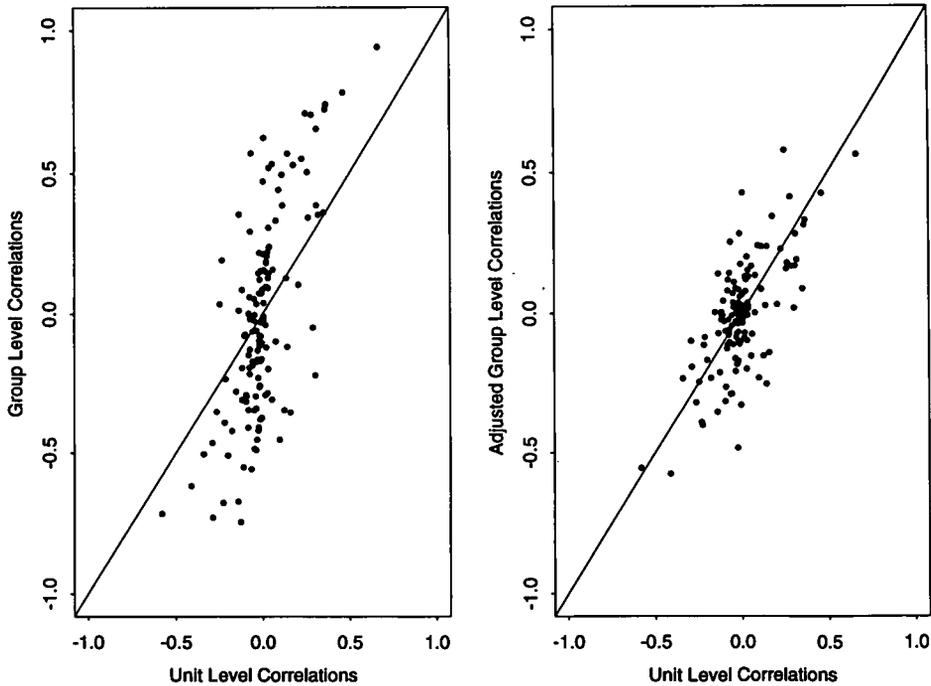


FIG. 4. Fig. 4a (left): Group-Level versus Unit-Level Correlations. Fig. 4b (right): Adjusted Group-Level versus Unit-Level Correlations.

Holt, and Tranmer 1996). The auxiliary variables must be those for which it is possible to obtain a reasonable estimate of the individual-level covariance matrix Σ_{zz} . For many variables it may be possible to obtain estimates of the individual-level variances. Potential auxiliary variables can then be identified from those variables with relatively large aggregation effects \bar{Q}_{aa} . Previous substantive research on these factors which are important determinants of where people live can also be a guide.

Occasionally it is possible to calculate a unit-level covariance matrix, S , for a range of sociodemographic and other variables, as well as the corresponding area-level covariance matrix \bar{S} , for example, from the population census. Steel and Holt (1996a) proposed using the eigenvectors of $S^{-1}\bar{S}$ to identify the key variables accounting for the aggregation effect in this set of variables. The variables identified in this way could then be considered as potential auxiliary variables to be used in adjusting the analysis of aggregate data for other sets of variables in the future. This method identifies those linear combinations of the variables with successively maximum aggregation effect, subject to being independent of each other. Steel and Holt (1996a) call these linear combinations "canonical grouping variables" (CGVs). The CGVs take into account the relationships between variables at the individual level and can help identify variables whose aggregation effects are mainly due to related variables. For example, applying this method to the seventeen variables listed in Table 1 supplemented by the housing variables listed in Table 2 suggests that much of the aggregation effect of the variables "0 Car Households" and " ≤ 0.5 persons/room" can be explained by the housing variables. For this reason, combined with the fact that individual-level data on the variables indicating car access and

TABLE 4
Components of Bias of Group-Level Variance

Variable	Unit-Level Variance	Bias	Bias due to Z	Residual Bias
Fulltime Employment	0.1170	0.8830	0.3896	0.4934
Unemployed	0.4402	0.5598	0.3936	0.1662
Other employment status	0.0893	0.9107	0.8516	0.0591
Nonwhite	0.1209	0.8791	0.8791	-0.0000
Female	0.9256	0.0744	0.3583	-0.2839
Aged 18-29	0.1087	0.8913	0.5545	0.3368
Aged 30-44	0.2192	0.7808	0.7241	0.0568
Aged 45-59	0.1674	0.8326	0.8326	0.0000
Aged 60 and over	0.0583	0.9417	0.9417	0.0000
HoH born U.K.	0.2234	0.7766	0.4274	0.3492
HoH born N.C.	0.2786	0.7214	0.6004	0.1211
Married	0.1602	0.8398	0.7078	0.1320
Limiting long-term illness	0.1381	0.8619	0.6938	0.1681
Migrant HoH	0.1106	0.8894	0.2558	0.6336
0 car households	0.2401	0.7599	0.3796	0.3803
Student \geq 18	0.0303	0.9697	0.8090	0.1607
\leq 0.5 persons/room	0.0358	0.9642	0.6039	0.3603

Source: Derived from 1991 Census SAS and SAR data; Crown Copyright.

density of occupancy are unlikely to be readily available in the years between population censuses, these variables were not included in the set of auxiliary variables considered in sections 3 and 4. See Table 4.

5. DISCUSSION

Statistical models that lead to a deeper understanding of aggregation effects and the cause of the ecological fallacy have been proposed for populations composed of geographic groups. The aggregation effects depend upon the sample sizes upon which the area means are based, the number of areas used in the analysis, and the strength of intra-area homogeneity on both variances and covariances for the variables of interest.

Auxiliary variables may be introduced that explain much of the intra-area homogeneity and hence the causes of the ecological fallacy. This leads to a decomposition of the aggregation bias into two components—one attributed to a set of grouping variables and the other a residual source of aggregation bias conditional on the grouping variables. If the grouping variables are powerful, the residual bias may be negligible and in many cases ought to be much smaller than the total aggregation bias.

With some additional information about the individual-level covariance matrix of the grouping variables, an adjustment is proposed that eliminates the first component of the aggregation bias. The empirical study suggests that this is a fruitful modification to ecological analyses.

LITERATURE CITED

- Amrhein, C. (1996). "Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulation." *Environment and Planning A*, forthcoming.
- Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Kluwen Academic Publishers.
- Blalock, H. M. (1964). *Casual Inference in Nonexperimental Research*. Chapel Hill, N.C.: University of North Carolina Press.

- (1979). "Measurement and Conceptualization Problems: The Major Obstacle to Integrating Theory and Research." *American Sociological Review* 44, 881–94.
- (1985). "Cross-Level Analysis." In *The Collection and Analysis of Community Data*, edited by J. B. Casterlin. ISI, World Fertility Survey.
- Clark, W. A. V., and K. L. Avery (1976). "The Effect of Data Aggregation in Statistical Analysis." *Geographical Analysis* 8, 428–38.
- Duncan, D. P., and B. Davis (1953). "An Alternative to Ecological Correlation." *American Sociological Review* 18, 665–66.
- Fotheringham, A. S., and D. W. S. Wong (1991). "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and Planning A* 23, 1025–44.
- Gehlke, C. E., and K. Biehl (1934). "Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association* 29, Supplement, 169–70.
- Goodman, L. A. (1959). "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64, 610–25.
- Hannan, M. T., and L. Burstein (1974). "Estimation from Grouped Observations." *American Sociological Review* 39, 374–92.
- Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953). *Sample Survey Methods and Theory*, Vol. 1, *Methods and Applications*. New York: John Wiley & Sons.
- Holt, D., D. G. Steel, and M. Tranmer (1996). "Area Homogeneity and the Modifiable Areal Unit Problem." *Geographical Systems*, forthcoming.
- Lichtman, A. J. (1974). "Correlation, Regression, and the Ecological Fallacy: A Critique." *Journal of Interdisciplinary History* 4, 417–33.
- Langbein, L. I., and A. J. Lichtman (1978). *Ecological Inference*. Beverley Hills, Calif.: Sage.
- Moellering, H., and W. Tobler (1972). "Geographical Variances." *Geographical Analysis* 4 (January), 34–50.
- Openshaw, S. (1984). "Ecological Fallacies and the Analysis of Areal Census Data." *Environment and Planning A* 6, 17–31.
- Openshaw, S., and P. J. Taylor (1979). "A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem." In *Statistical Applications in the Spatial Sciences*, edited by N. Wrigley, pp. 127–44. London.
- Robinson, W. S. (1950). "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15, 351–57.
- Smith, K. W. (1977). "Another Look at the Clustering Perspective in Aggregation Problems." *Sociological Methods and Research* 5, 289–316.
- Steel, D. (1985). "Statistical Analysis of Populations with Group Structure." Ph.D. thesis, Department of Social Statistics, University of Southampton.
- Steel, D., and D. Holt (1996a). "Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisited." *International Statistical Review*, forthcoming.
- (1996b). "Rules of Random Aggregation." *Environment and Planning*, forthcoming.
- Steel, D., D. Holt, and M. Tranmer (1996). "Making Unit-Level Inferences from Aggregated Data." *Survey Methodology*, forthcoming.
- Yule, U., and M. S. Kendall (1950). *An Introduction to the Theory of Statistics*. London: Charles Griffin.